Development and Initial Testing of a Parallel Ensemble Kalman Filter for the Poseidon Isopycnal Ocean General Circulation Model

Christian L. Keppenne
Science Applications International Corporation
4600 Powder Mill Road
Beltsville, Maryland 20705

Michele M. Rienecker
NASA Seasonal-to-Interannual Prediction Project
Code 971, Laboratory for Hydrospheric Processes
Goddard Space Flight Center, Greenbelt, Maryland 20771

Submitted to *Monthly Weather Review*

August 2001

Corresponding author address:

Christian L. Keppenne
Mail Code 971, NASA Goddard Space Flight Center, Greenbelt, Maryland 20771
e-mail: clk@janus.gsfc.nasa.gov

Abstract

A multivariate ensemble Kalman filter (MvEnKF) implemented on a massively parallel computer architecture has been developed for the Poseidon ocean circulation model and tested with a Pacific Basin model configuration. There are about two million prognostic state-vector variables. Parallelism for the data assimilation step is achieved by regionalization of the background-error covariances that are calculated from the phase-space distribution of the ensemble. Each processing element (PE) collects elements of a matrix measurement functional from nearby PEs. To avoid the introduction of spurious long-range covariances associated with finite ensemble sizes, the background-error covariances are given compact support by means of a Hadamard (element by element) product with a three-dimensional canonical correlation function.

The methodology and the MvEnKF implementation are discussed. To verify the proper functioning of the algorithms, results from an initial experiment with *in situ* temperature data are presented. Furthermore, it is shown that the regionalization of the background covariances has a negligible impact on the quality of the analyses.

The parallel algorithm is very efficient for large numbers of observations. On a platform with distributed memory, individual-PE memory, rather than speed, dictates how large an ensemble can be used in practice.

1 Introduction

a. Background and motivation

Many of the early advances in ocean data assimilation have emerged from practical applications in the tropical Pacific. These applications have been driven by the need to initialize the ocean state for coupled atmosphere-ocean forecasts of the El Niño-Southern Oscillation (ENSO) phenomenon. In addition, hindcast estimates of the ocean state have been useful in diagnosing the evolution of El Niño. Over much of the world's oceans, large-scale assimilation is facilitated by the availability of satellite altimetry because of the sparsity of *in situ* data. However, in the tropical Pacific, the ocean observing system was vastly improved by the deployment of the Tropical Atmosphere-Ocean (TAO) array of moored buoys (e.g., McPhaden *et al.* 1998) to support seasonal-to-interannual (SI) climate studies and prediction. One of the major successes of the Tropical Ocean Global Atmosphere (TOGA) program was the emergence of coupled physical models (as opposed to statistical models) with some prediction skill (*e.g.,* Chen *et al.* 1995; Ji *et al.* 1996).

Recently, the NASA Seasonal-to-Interannual Prediction Project (NSIPP) has been established to further the utilization of satellite observations for prediction of short term climate phenomena. NSIPP undertakes routine forecasts in a research framework with global coupled ocean-atmosphere-land surface models. The initial implementation has used an ocean analysis system employing a simple assimilation methodology—a univariate optimal interpolation (UOI, *e.g.,* Troccoli *et al.* 2001)—with the Poseidon isopycnal ocean general circulation model (OGCM, Schopf and Loughe 1995; Konchady *et al.* 1998; Yang *et al.* 1999). Like several other ocean data assimilation systems currently in use at other institutions (*e.g.*, Ji and Leetma 1997), it is

based on the assumption that the forecast-error covariances are approximately Gaussian and that the covariances between the temperature-field errors and the salinity-field and current-field errors are negligible.

Largely due to the high-resolution coverage and accuracy of the TAO measurements, the UOI is effective in improving surface and sub-surface temperature field estimates in the equatorial region in comparison with the estimates obtained without temperature assimilation. As a result, its introduction into the NSIPP coupled forecasting system has resulted in significant improvements in the coupled model's hindcast skill of Niño-3 temperature anomalies.

The UOI has the advantage of being inexpensive in terms of computing resources. Nevertheless, it suffers from three major shortcomings. The first shortcoming is that it can only be used to assimilate measurements of a model prognostic variable. The UOI's second shortcoming is that it does not use any statistical information about the expected inhomogeneous distribution of model errors. The third shortcoming is that it is based on a steady state error-covariance model which gives the same weight to a unit innovation regardless of how accurate the ocean-state estimate has become as a result of previous analyses. Directly linked to this shortcoming is the failure to provide time-dependent estimates of the model errors.

In response to the first two shortcomings, a parallel multivariate OI (MvOI) system has been implemented. The MvOI uses steady state estimates of the model-error statistics computed from ensemble runs of the OGCM in the presence of stochastic atmospheric forcing from Monte Carlo simulations (Borovikov and Rienecker 2001). Yet, the MvOI cannot adjust to dynamically

evolving error statistics. A parallel multivariate ensemble Kalman filter (MvEnKF) has been developed to address this shortcoming. This paper discusses its design, implementation and initial testing.

b. Overview of the ensemble Kalman filter

Although the Kalman filter (Kalman 1960) and its generalization to nonlinear systems, the extended Kalman filter, are statistically optimal sequential estimation procedures that minimize error variance (Daley 1991; Ghil and Malanotte-Rizzoli 1991; Bennett 1992; Robinson *et al.* 1998), they cannot be used in the context of a high-resolution ocean or atmospheric model because of the prohibitive cost of time stepping the model-error covariance matrix when the model has more than a few thousand state variables. Therefore, reduced-rank (*e.g.*, Cane *et al.* 1996, Verlaan and Heemink 1997) and asymptotic (e.g., Fukumori and Malanotte-Rizzoli 1995) Kalman filters have been proposed. Evensen (1994) introduced the ensemble Kalman filter (EnKF) as a Monte Carlo-based alternative to the traditional Kalman filter. In the EnKF, an ensemble of model trajectories is integrated and the statistics of the ensemble are used to estimate the model errors. Closely related to the EnKF are the singular evolutive extended Kalman filter (Pham *et al.* 1998) and the error-subspace statistical estimation algorithms described in Lermusiaux and Robinson (1999).

Evensen (1994) compared the EnKF to the extended Kalman filter in twin assimilation experiments involving a two-layer quasigeostrophic (QG) ocean model on a square 17 $\times$ 17 grid. Evensen and van Leeuwen (1996) used the EnKF to process GEOSAT altimeter data into a two-layer, regional QG model of the Agulhas current on a 51 $\times$ 65 grid. Houtekamer and

Mitchell (1998) and Mitchell and Houtekamer (2000) used the EnKF in identical-twin experiments involving a three-level, spectral QG model at triangular truncation T21 and parameterized model errors.

Keppenne (2000, hereafter K00) conducted twin experiments with a parallel MvEnKF algorithm implemented for a two-layer, spectral, T100 primitive equation model with parameterized model errors. With about 2×10^5 model variables, the state-vector size was small enough in this application to justify a parallelization scheme in which each ensemble member resides in the memory of a separate CRAY T3E processor (hereafter processing element: PE). To parallelize the analysis, K00's algorithm transposes the ensemble across PEs at analysis time, so that each PE ends up processing data from a sub-region of the model domain. The influence of each observation is weighted according to the distance between that observation and the center of each PE region.

To filter out noise associated with small ensemble sizes, Houtekamer and Mitchell (2001) developed a parallel EnKF analysis algorithm that applies a Hadamard (element by element) product (*e.g.*, Horn and Johnson 1991) of a correlation function having local compact support with the background-error covariances. They tested this analysis scheme on a 128×64 Gaussian grid corresponding to a three-level QG model using randomly generated ensembles of first-guess fields computed ahead of time. The benefits of constraining the covariances between ensemble members using a Hadamard product with a locally supported correlation function has also been investigated by Hamill and Snyder (2000) in the context of an intermediate QG atmospheric model.

In this paper, we build upon the contributions made by each of the above-mentioned studies to implement a parallel MvEnKF for the Poseidon OGCM. Initial tests are undertaken with a 20-layer, Pacific basin configuration of the model with about two million state variables. The system noise is accounted for in a manner similar to that used in K00, by including a stochastic component in the forcing fields. Following Houtekamer and Mitchell (2001), an element-by-element product with an idealized three-dimensional compactly supported correlation function is used to remove spurious long-range signals from the background-error covariances.

c. *Organization of the following Sections*

The remainder of this paper is concerned with the parallel MvEnKF design for the Poseidon OGCM. The model is briefly discussed in Section 2 and the algorithms are presented in Section 3. The scalability of the algorithms and the effect of distributing the analysis calculations between PEs are discussed in Section 4, where an initial test of the MvEnKF is conducted in the context of TAO-temperature data assimilation. Section 5 contains a summary.

A complete description of the algorithms and of the multivariate data assimilation system is available under the form of a NASA technical report (Keppenne and Rienecker 2001a, hereafter KR01a). The application of the MvEnKF to the assimilation of altimeter data into Poseidon is discussed in Keppenne and Rienecker (2001b, hereafter KR01b). Its application to the assimilation of TAO temperature data including the impact of the assimilation on the model currents, salinity and sea surface height (SSH) are the focus of another article in which the

MvEnKF is also compared to the UOI currently used to initialize the NSIPP SI forecasts (Keppenne and Rienecker 2001c, hereafter KR01c).

2 The Poseidon parallel ocean model

a. Model summary

The Poseidon model (Schopf and Loughe, 1995) is a finite-difference reduced-gravity ocean model which uses a generalized vertical coordinate designed to represent a turbulent, well-mixed surface layer and nearly isopycnal deeper layers. Poseidon has been documented and validated in hindcast studies of El Niño (Schopf and Loughe 1995) and has since been updated to include prognostic salinity (*e.g.*, Yang *et al.* 1999). More recently, the model has been used in a numerical study of the surface heat balance along the equator (Borovikov *et al.* 2001) and in an examination of ENSO and its mechanisms during the 1990s (Yuan *et al.* 2001).

Explicit detail of the model, its vertical coordinate representation and its discretization are provided in Schopf and Loughe (1995). The prognostic variables are layer thickness, $h(\lambda, \theta, \zeta, t)$, temperature, $T(\lambda, \theta, \zeta, t)$, salinity, $S(\lambda, \theta, \zeta, t)$, and the zonal and meridional current components, $u(\lambda, \theta, \zeta, t)$ and $v(\lambda, \theta, \zeta, t)$, where λ is longitude, θ latitude, t time and ζ is a generalized vertical coordinate which is 0 at the surface and increments by 1 between successive layer interfaces.

Following Pacanowski and Philander (1981), vertical mixing is parameterized through a Richardson number-dependent mixing scheme implemented implicitly. An explicit mixed layer is included with a mixed layer entrainment parameterization following Niiler and Kraus (1977).

A time-splitting integration scheme is used whereby the hydrodynamics are done with a short time step (15 minutes), but the vertical diffusion, convective adjustment and filtering are done with coarser time resolution (half-daily).

b. Model setup

The version of Poseidon used here has been parallelized as in Konchady *et al.* (1998) using the same message-passing protocol and 2D horizontal domain decomposition used by Schaffer and Suarez (1998) for the NSIPP-1 atmospheric general circulation model.

The experiments of Section 4 use a 20-layer Pacific basin version of the parallel model with uniform 1º zonal resolution. The meridional resolution varies between 1/3º at the equator and 1º in the extratropics. A solid boundary is imposed at 45º south. There, a no-slip condition is used for the currents and a no-flux condition is used for mass, heat and salinity.

There are $173 \times 164 \times 20$ grid boxes, of which 28% are situated over land, resulting in a total of 2.0422×10^6 individual prognostic variables. A 16×16 PE lattice is used as shown in Figure 1. The PEs located over land are virtual PEs which do not take part in the ensemble integrations and analyses.

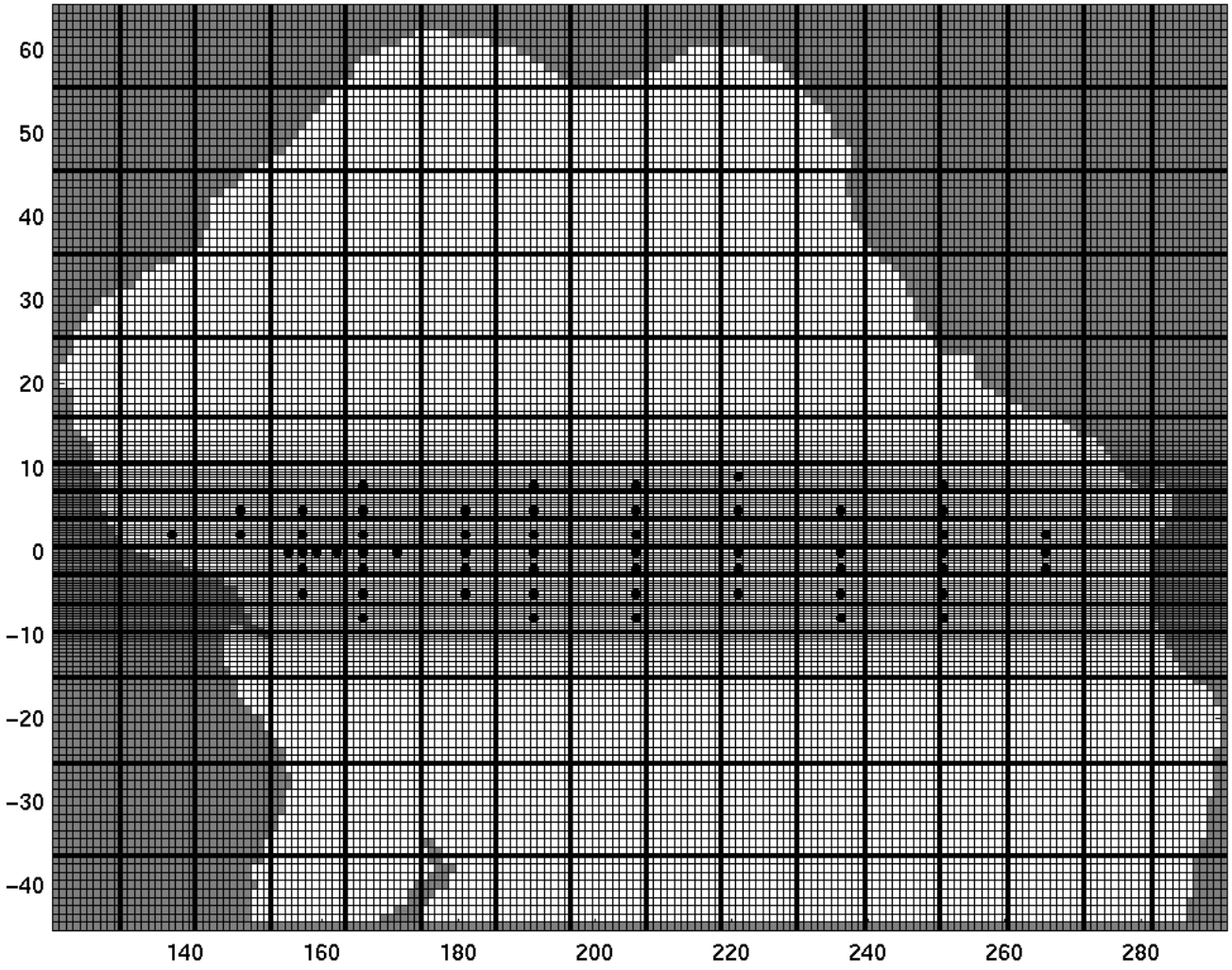

Figure 1. Horizontal domain decomposition for the Pacific model. The thin lines delineate grid cells. The thick lines correspond to the boundaries of each PE box's PE-private area on the 16×16 PE lattice. Each dark circle corresponds to a TAO mooring.

Figure 2 illustrates the horizontal setup for one PE box. Locally within the box, the grid cells are numbered $1 \le i \le I$, zonally and $1 \le j \le J$, meridionally, from the box's lower-left, southwest corner. In order to minimize the communication overhead in the horizontal differencing of the model equations, the PE boxes overlap. The overlapping regions, called halo regions, have width $i_1 - 1$ to the West, $I - i_2$ to the East, $j_1 - 1$ to the South and $J - j_2$ to the North. The PE-private regions are thus defined by $i_1 \le i \le i_2$ and $j_1 \le j \le j_2$.

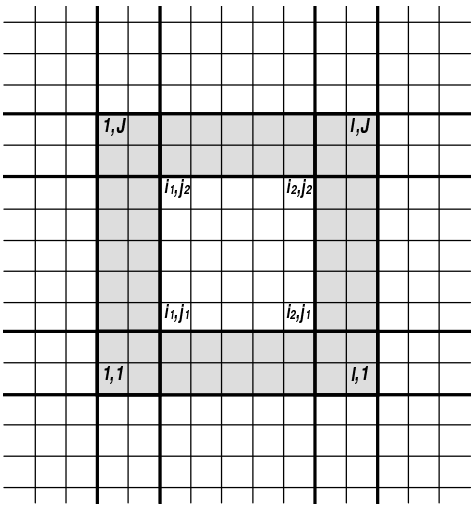

Figure 2. Schematic setup for one PE. The halo regions are colored gray. The thin lines delineate grid cells. The thick lines delimit the halo regions and PE boundaries. In this example, $I = J = 9$, $i_1 = j_1 = 3$ and $i_2 = j_2 = 7$.

3 Assimilation methodology

a. Horizontal domain decomposition

Since the version of Poseidon used here is parallelized, the same domain decomposition used to run the model can be used in the analyses, provided the background error-covariance matrix, $\boldsymbol{P}^f$, is locally approximated. This simplification avoids costly ensemble transpositions across PEs. Thus, the ensemble is distributed so that the memory of each PE contains the same elements of each ensemble member's state vector. These elements correspond to every variable contained within the PE boxes, the PE-private portions of which are visible in Figure 1. This decomposition is used for the ensemble integrations as well as for the analyses.

b. Assimilation on geopotential surfaces

The temperature measurements from each TAO mooring are recorded at specific depths which are fairly consistent between moorings. Since Poseidon uses an isopycnal vertical coordinate, the model fields must be interpolated to the latitude, longitude and depth of each observation. When the UOI was implemented, the choice was made to treat the temperature observations in the usual (λ, θ, z) coordinate system in light of the absence of corresponding salinity observations. To maintain compatibility with the UOI which interpolates model fields vertically to a series of pre-specified depths (hereafter levels) prior to each analysis, the same approach is used here and the background covariances are calculated on levels rather than on layers. Therefore, the T, S, u and v fields are converted from isopycnals to levels and the analysis increments are calculated on the levels before being mapped back to the isopycnals. Sixteen levels are used in Section 4 and in KR01b-c.

The above scheme results in only *T*, *S*, *u* and *v* being updated. The layer thicknesses, *h*, are left unchanged by the assimilation. The procedure allows the model to dynamically recalculate *h* from the new density distribution and the target interface buoyancies, as it does at every time step (see Schopf and Loughe 1995).

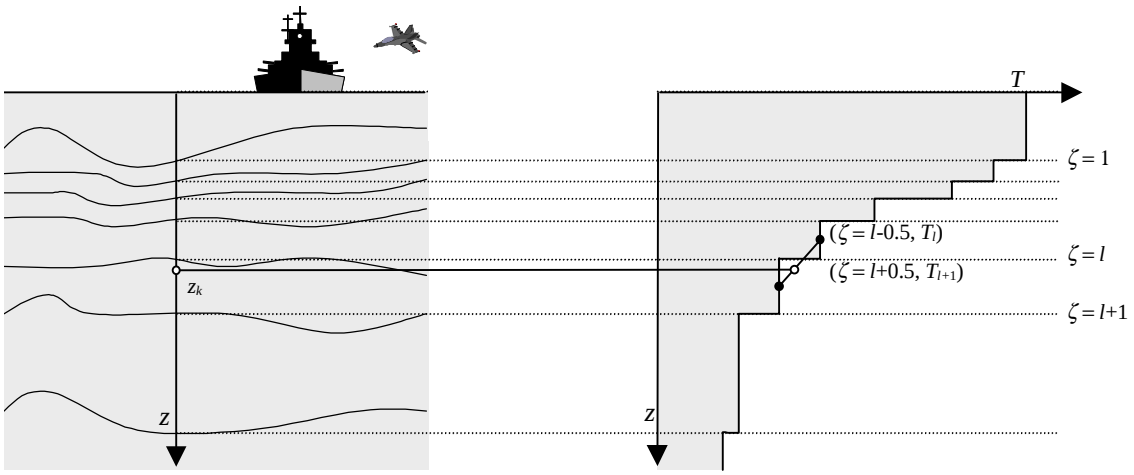

Figure 3. Mapping of the model temperature field to a specified level, $z = z_k$. Within the current grid cell, z_k is contained between the layer interfaces $\zeta = l$ and $\zeta = l+1$. In the model discretization, only the layer-average temperature matters. Yet, to avoid ambiguities when more than one specified level pass through the same layer in the grid cell, the field is interpolated linearly as shown.

Since only the layer-average value of *T, S, u* and *v* in each vertical grid box (*i, j, l*) within grid cell (*i, j*) appear in the model equations, the mapping from isopycnals to levels could be made by assigning to a given field at (λ_{ij}, θ_{ij}, z_k).the value of the same field at (λ_{ij}, θ_{ij}, *l*). However, if the

mapping were performed in this manner, ambiguities would arise when several levels pass through the same layer at (λ_{ij}, θ_{ij}). A possible consequence is the singularity of the analysis equations of Section 3e in the (λ, θ, z) coordinate system. To avoid this problem, the mapping is made as though the vertical variations of the field were piecewise linear, with the discontinuities in the slope occurring in the middle of the layers. This is illustrated in Figure 3 for the temperature field.

c. Ensemble size

With the MvEnKF, PE memory imposes constraints on both the domain decomposition and the ensemble size. The Pacific basin version of Poseidon is typically run on 64 PEs. The goal is for the MvEnKF runs to be done on a few times as many PEs. In this study, 256 PEs are used and the memory available on these PEs imposes a limit of about 40 ensemble members on the platform currently used for the production forecasts (1024-PE CRAY T3E-600 with 128MB local RAM per PE). Encouraging results have been obtained with comparably sized ensembles by Mitchell and Houtekamer (2000) with a three-level QG model and by K00 with a two-layer shallow water model. The issue of whether the MvEnKF perform as well as or better than the UOI with even as few as 40 ensemble members is touched upon in Section 4a and examined in depth in KR01c.

d Decomposition of analysis between PEs

The small ensemble size introduces the need to filter out spurious long-range correlations when the background covariances are computed. Following Houtekamer and Mitchell (2000) and a suggestion by Gaspari and Cohn (1999), this filtering is achieved through a Hadamard product

(*i.e.* $\boldsymbol{A} \bullet \boldsymbol{B}$ such that $\{\boldsymbol{A} \bullet \boldsymbol{B}\}_{ij} = \boldsymbol{A}_{ij} \boldsymbol{B}_{ij}$) of the error-covariance matrices with a local compactly supported correlation function. This function is the product of a horizontal correlation function, $C_h(r_h^{(12)})$, $r_h^{(12)} = [(\lambda_2 - \lambda_1)^2/l_\lambda + (\theta_2 - \theta_1)^2/l_\theta]^{0.5}$, and a vertical correlation function, $C_v(r_v^{(12)})$, $r_v^{(12)} = |z_2 - z_1|/l_z$, where $(\lambda_i, \theta_i, z_i)$ are the coordinates of point i. In this study, $C_h = C_v = C_0$, where C_0 is defined by (4.10) of Gaspari and Cohn (1999). The normalization is such that $C_0(r) = 0, \quad r \geq 2$. The correlation scales used in Section 4 and in KR01c for the assimilation of TAO temperature data are $l_\lambda = 30^o$, $l_\theta = 15^o$ and $l_z = 500$m. Shorter correlation scales give better results when gridded altimeter data are assimilated (KR01b).

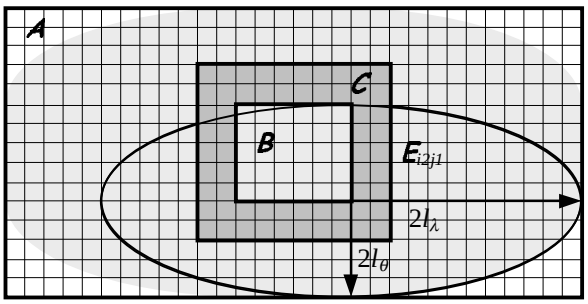

Figure 4. Domain decomposition for the analysis. The outer rectangle delimits the area, ***A***, from which the data assimilated on one PE are collected. The innermost rectangle depicts the boundary of the PE-private area, ***B***. The ellipse delimits the influence region of the PE-private area's southeastern corner cell, (i_2, j_1). The shaded area contains the ellipses for all grid cells, (i, j), contained in ***B***. The region ***C*** contains all the PE's grid cells including the halo regions.

Although the TAO temperature data assimilated in Section 4 are sufficiently few (about 600 at each analysis) for each PE to process them all, an approach whereby each PE processes data from a sub-region of the model domain is used. When more numerous data are assimilated, such as in KR01b, the regionalization becomes a necessity.

Besides the obvious efficiency gain in a parallel environment, another justification for decomposing the analysis is that the compactly supported background covariances result in the data that directly (*i.e.*, through the measurement operator) influence the state variables within each grid cell being contained within an ellipse with semi axes $2l_\lambda$ and $2l_\theta$. Taking advantage of this fact, the region from which the observations assimilated on each PE are collected is chosen to be the smallest rectangle, with sides $\lambda_{i2j1} - \lambda_{i1j1} + 4l_\lambda$ and $\theta_{i1j2} - \theta_{i1j1} + 4l_\theta$, containing all the ellipses that correspond to the PE-private grid cells of this PE. This is illustrated in Figure 4.

e. Analysis procedure

Without the Hadamard product of the background-error covariances with the compactly supported correlation function, the EnKF analysis can be written as

$$\boldsymbol{y}_i = \boldsymbol{\Xi}(\boldsymbol{x}_i^f - \langle \boldsymbol{x} \rangle^f), \qquad (1a)$$

$$\boldsymbol{l}_i = \boldsymbol{\mathcal{L}}(\boldsymbol{y}_i + \langle \boldsymbol{x} \rangle^f) - \boldsymbol{\mathcal{L}}(\langle \boldsymbol{x} \rangle^f), \qquad (1b)$$

$$\boldsymbol{Y} = \{\boldsymbol{y}_1, \cdots, \boldsymbol{y}_m\}, \quad \boldsymbol{L} = \{\boldsymbol{l}_1, \cdots, \boldsymbol{l}_m\},$$

$$\left[\boldsymbol{L}\boldsymbol{L}^T + \boldsymbol{W}\right]\boldsymbol{b}_i = \boldsymbol{d} - \boldsymbol{\mathcal{L}}(\boldsymbol{y}_i + \langle \boldsymbol{x} \rangle^f) + \boldsymbol{e}_i, \qquad (1c)$$

$$\boldsymbol{x}_i^a = \boldsymbol{x}_i^f + \boldsymbol{Y}\boldsymbol{L}^T\boldsymbol{b}_i. \qquad (1d)$$

In (1) and throughout this discussion, uppercase boldface symbols represent matrices, lowercase boldface symbols represent vectors and lowercase regular (*i.e.*, not bold) symbols denote scalar variables. The vector, $\boldsymbol{d}$ $(n_d \times 1)$ contains n_d observations, $\boldsymbol{x}_i$ $(n_x \times 1)$, $1 \le i \le m$, is the ith ensemble state vector of length n_x and m stands for the ensemble size. The superscripts ^{a} and ^{f} refer to the analyzed state and the forecast, respectively, Ξ is a smoothing operator (Section 3h) and < > denotes an ensemble average. The vectors $\boldsymbol{y}_i$ $(n_x \times 1)$ and $\boldsymbol{l}_i$ $(n_d \times 1)$ are columns of the matrices $\boldsymbol{Y}$ $(n_x \times m)$ and $\boldsymbol{L}$ $(n_d \times m)$ respectively, and $\boldsymbol{\mathcal{L}}(\boldsymbol{x})$ is a measurement operator which relates the state vector to the observations. Matrix $\boldsymbol{W}$ $(n_d \times n_d)$ is the observation-error covariance matrix. It includes measurement errors as well as representation errors. The representer matrix, $\boldsymbol{R} = \boldsymbol{L}\,\boldsymbol{L}^T$, maps the background-error covariance matrix, $\boldsymbol{P}^f$ $(n_x \times n_x)$, to the error subspace of the measurements. The elements of $\boldsymbol{b}_i$ are the representer-function amplitudes used to update $\boldsymbol{x}_i$.

The $n_d \times 1$ vector, $\boldsymbol{z}_i = \boldsymbol{d} - \boldsymbol{\mathcal{L}}(\boldsymbol{y}_i + \langle \boldsymbol{x} \rangle^f) + \boldsymbol{e}_i$ in (1c), contains the innovations with respect to the ith ensemble member. Prior to their calculation, Ξ is applied to smooth $\boldsymbol{x}_i$. Following Burgers *et al.* (1998), $\boldsymbol{e}_i$ is a random perturbation chosen such that $\langle \boldsymbol{e}_i \rangle = 0$ and $\left\langle \boldsymbol{e}_i \boldsymbol{e}_i^{\,T} \right\rangle = \boldsymbol{W}$. Its role is to maintain the influence of observation uncertainty in the error covariances estimated directly from the ensemble so that these covariances are consistent with the theoretical estimates. Its inclusion helps prevent the ensemble from collapsing resulting in a systematic error underestimation.

When $\boldsymbol{\mathcal{L}}(\boldsymbol{x})$ is a linear operator and Ξ is an identity mapping, (1) simplifies to the usual Kalman filter analysis equations (*e.g.*, Gelb, 1974) applied to update each ensemble member in turn.

When the Hadamard products with the compactly supported correlation function are introduced and when the subscript ranges are explicitly written down, (1c) and (1d) are replaced by

$$c_{pq} = c_{qp} = C_h(r_h^{(pq)})\, C_v(r_v^{(pq)}), \qquad 1 \le p \le n_d, \quad 1 \le q \le n_d, \qquad (2a)$$

$$\left[\boldsymbol{C} \bullet \boldsymbol{L}\boldsymbol{L}^T + \boldsymbol{W}\right]\boldsymbol{b}_i = \boldsymbol{d} - \boldsymbol{\mathcal{L}}(\boldsymbol{y}_i + \langle \boldsymbol{x} \rangle^f) + \boldsymbol{e}_i, \qquad 1 \le i \le m, \qquad (2b)$$

$$\eta_{kp} = C_h(r_h^{(kp)})\, C_v(r_v^{(kp)}), \qquad 1 \le k \le n_{box}, \quad 1 \le p \le n_d, \qquad (2c)$$

$$\left.\begin{aligned} \gamma_{ik} &= \boldsymbol{L}^T \boldsymbol{b}_i \bullet \boldsymbol{\eta}_k, \\ x_{ik}^a &= x_{ik}^f + \boldsymbol{y}_k \circ \boldsymbol{\gamma}_{ik}, \end{aligned}\right\} \quad 1 \le i \le m, \quad 1 \le k \le n_{box}, \qquad \begin{matrix}(2d)\\(2e)\end{matrix}$$

where $\circ$ and $\bullet$ refer respectively to the inner product of two vectors and to the Hadamard product of two matrices, and $\boldsymbol{C}$ ($n_d \times n_d$) is a compactly supported correlation matrix whose elements are defined by (2a), where the indices p and q refer to the data w_p and w_q. The components of the $n_d \times 1$ vector $\boldsymbol{\eta}_k$ defined by (2c) contain idealized correlations between the (λ, θ, z) coordinates, of grid box k and the coordinates of each measurement. To simplify the notation, only one subscript is used to identify the grid box. The index, $1 \le k \le n_{box}$, thus loops over the three dimensions of the (λ, θ, z) coordinate system. The $m \times 1$ vector, $\boldsymbol{y}_k = \{y_{1k}, \cdots, y_{mk}\}$, contains smoothed deviations from the ensemble mean of the m ensemble state vectors in the kth grid box. It is thus a single row of matrix $\boldsymbol{Y}$. With the MvEnKF, y_{ik} actually has four components, *i.e.*,

$$y_{ik} = \Xi(\{T,S,u,v\}_{ik} - \{\langle T\rangle,\langle S\rangle,\langle u\rangle,\langle v\rangle\}_k).$$

The $m \times 1$ vector, $\boldsymbol{\gamma}_{ik}$, contains the weights with which the elements of $\boldsymbol{y}_k$ in the *k*th grid box, $\{T,S,u,v\}_{ik}$, are combined to update the *i*th ensemble member. In each grid box, the analysis update, (2c-e), involves *m* matrix-vector multiplication of $\boldsymbol{L}^T$ by $\boldsymbol{b}_i \bullet \boldsymbol{\eta}_k$ (2d). If the analysis calculations were not distributed as explained in Section 3g, or if the observations allowed to influence the variables of each grid box were not limited to a sub-region of the entire domain as a result of imposing compactly supported background covariances, these multiplications would be very costly. For the Poseidon model distributed across 256 PEs, they correspond to a tolerable fraction of the total cost of the MvEnKF. For example, when TAO temperature data are assimilated every five days as in Section 4, the ensemble integration takes about 1100 seconds per analysis cycle while the analysis takes about 380 seconds. Of these, about 270 seconds are spent in the matrix-vector products of (2d).

f. System-noise representation

The theory of the Kalman filter (*e.g.*, Gelb, 1974) assumes that the first- and second-order statistics of the errors in the model and external forcing are known. Higher-order statistics are neglected. Let the evolution of the true state be represented by

$$\frac{\partial \boldsymbol{x}}{\partial t} = \boldsymbol{F}(\boldsymbol{x},t) + \boldsymbol{\xi}(\boldsymbol{x},t), \qquad (3)$$

where $\boldsymbol{\xi}$ combines the model errors and forcing errors, and is commonly known as system noise or process noise. The vector $\boldsymbol{F}$ $(n_x \times 1)$ contains the right hand sides of the model partial differential equations which includes the model hydrodynamics, physics and forcing. It is assumed that the model and forcing are unbiased, *i.e.* $\langle \boldsymbol{\xi}(\boldsymbol{x},t) \rangle = 0$, and that the $\boldsymbol{\xi}$ vectors are uncorrelated in time:

$$\langle \boldsymbol{\xi}(\boldsymbol{x}_k,t_k)\, \boldsymbol{\xi}(\boldsymbol{x}_l,t_l) \rangle = \boldsymbol{\Gamma}(\boldsymbol{x}_k,\boldsymbol{x}_l)\, \delta(t_k - t_l), \qquad (4)$$

where the system-noise covariance matrix, $\boldsymbol{\Gamma}$, is assumed known. Of course, the unbiased assumption is rarely correct in practice. This is especially true with ocean models in which the thermocline layer is usually too diffuse. In an effort to account for the model bias, an algorithm, derived from Dee and Da Silva (1998), to estimate and correct systematic model errors has recently been implemented into the MvEnKF code. However, it is not used in this study.

In meteorological and oceanographic data assimilation, the statistics of $\boldsymbol{\xi}$ are generally unknown and are the object of parameterization. Adaptive Kalman filters that simultaneously estimate the state and system-noise statistics have been developed. Blanchet and Frankignoul (1997) summarize and compare several adaptive filtering algorithms. In practice, the prohibitive cost of the adaptive filters has limited their application in meteorology and oceanoraphy.

Motivated by the current lack of information about the model-error statistics, the system-noise is represented solely by modeling the errors in the surface wind stress and heat flux forcing. A

system-noise representation in which not only the forcing errors but also the model errors are parameterized is in development.

Because of the focus on SI variability, the forcing errors (uncertainties) are modeled on those time scales, with each ensemble member being forced by a monthly mean perturbation of the monthly mean basic state. The basic state is the superposition of the climatological seasonal cycle with interannual anomalies. The climatology is provided by Special Sensor Microwave Imager (Atlas *et al.* 1996) winds and Earth Radiation Budget Experiment heat flux data. The interannual anomalies are obtained by integrating the atmospheric model over observed SST data (Reynolds and Smith 1994). The perturbations applied are due entirely to internal atmospheric chaos and are generated by starting the atmospheric integrations at different times. By using the same SST, each member of the atmospheric ensemble used to force the ocean ensemble has the same SI phase. The spread of the atmospheric ensemble is meant to be representative of the uncertainty of the forcing products used to force the model in non-ensemble runs.

g. Parallel algorithm

1) Preliminaries

This section discusses what steps are involved in the parallel MvEnKF analyses from the point of view of one PE, hereafter referred to as the current PE. This overview starts after the current PE has obtained the observations, $\boldsymbol{d}^b$, made within its PE-private region ($\boldsymbol{\mathcal{B}}$ in Fig. 4). The position of each PE on the lattice is stored in the $n_{PE} \times m_{PE}$ array $\boldsymbol{PE}$, where n_{PE} and m_{PE} are the number of PEs along the zonal and meridional directions, respectively. In Section 4 and in KR01b-c, $n_{PE} = m_{PE} = 16$. The total number of PEs is N_{PE}. Every PE has a copy of $\boldsymbol{PE}$.

All information exchanges between PEs use message-passing functions from the Goddard Earth Modeling System (GEMS, Schaffer and Suarez 1998) library. The GEMS functions provide a high-level, object oriented interface to the CRAY native SHMEM (shared memory) communication library.

The analyses rely principally on two GEMS functions which are mentioned here in template form in order to simplify the discussion. The first function, **`pe_collect(...)`**, is used to collect data from either the entire ***PE*** array or from the row or column of ***PE*** which contains the current PE. The second function, **`halo(...)`**, updates its array argument in the halo regions of each PE (gray areas in Fig. 2), after each PE has modified its PE-private elements of this array corresponding to the inner rectangle in Figure 2.

2) Algorithm

- Step 1: Vertical interpolation of the *T*, *S*, *u* and *v* fields from the model layers to the analysis levels as explained in Section 3b.
- Step 2: Calculation of the anomalies with respect to the ensemble mean over the entire domain of the current PE (area $\boldsymbol{\mathcal{C}}$ in Fig. 4), $\boldsymbol{x}_i^{\boldsymbol{c}f} - \left\langle \boldsymbol{x}^{\boldsymbol{c}} \right\rangle^f, 1 \le i \le m.$
- Step 3: Calculation of $\boldsymbol{y}_i^{\boldsymbol{c}}$, the current PE's portion of $\boldsymbol{y}_i$ in (1a) . Prior to each zonal application of the smoothing operator, Ξ, a call to **`pe_collect( )`** is used to collect, from the PEs listed in column *jc* of ***PE***, the state elements required to run a recursive filter

(Section 3h). The same holds for each meridional application of the filter, where row *ic* of ***PE*** is now involved.

- Step 4: Identification of the PE-private data required by the other PEs. First, **`pe_collect( )`** is used to collect the longitudes and latitudes of each PE's southwestern, southeastern, northwestern and northeastern corner grid cells. Using this information, the current PE calculates for each (*i*, *j*) pair which elements of its $\boldsymbol{d}^{\boldsymbol{b}}$ fall inside the rectangle, $\mathcal{A}_{ij}$, which is the region from which PE_{ij} will need to collect data (Fig. 4). The indices of the relevant elements of $\boldsymbol{d}^{\boldsymbol{b}}$ are stored in an array, $\boldsymbol{k}_{ij}$.

- Step 5: Evaluation of the measurement operator. The current PE calculates a $n_d^{\boldsymbol{b}} \times m$ matrix, $\boldsymbol{L}^{\boldsymbol{b}}$, where $n_d^{\boldsymbol{b}}$ is the number of observations contained in its PE-private region. The element at the intersection of the *p*th row and *i*th column of $\boldsymbol{L}^{\boldsymbol{b}}$ is

$$L_{pi}^{\boldsymbol{b}} = \mathcal{L}^p(\boldsymbol{y}_i + \langle \boldsymbol{x} \rangle^f) - \mathcal{L}^p(\langle \boldsymbol{x} \rangle^f),$$

where $\mathcal{L}^p$ is an interpolation operator which maps its argument to the location of $d_p^{\boldsymbol{b}}$, the *p*th PE-private observation on the current PE (KR01a).

- Step 6: Calculation of $\boldsymbol{z}^{\boldsymbol{b}}$, the innovations with respect to the ensemble mean for the current PE's private region. The innovation corresponding to $d_p^{\boldsymbol{b}}$ is

$$z_p^{\boldsymbol{b}} = d_p^{\boldsymbol{b}} - \mathcal{L}^p(\langle \boldsymbol{x}^c \rangle^f).$$

- Step 7: Gathering of $\boldsymbol{L}^a$, a $n_d^a \times m$ matrix analogous to $\boldsymbol{L}^b$, but corresponding to the n_d^a measurements made within area $\boldsymbol{A}$ (Fig. 4), using the information recorded in the $\boldsymbol{k}_{ij}$ arrays. The function **pe_collect( )** is called N_{PE} times. Each call results in a different PE completing the collection of its version of $\boldsymbol{L}^a$.

- Step 8: Collection of the innovations, $\boldsymbol{z}^a$, required by each PE. As for gathering $\boldsymbol{L}^a$, **pe_collect( )** is called N_{PE} times. Each PE passes to **pe_collect( )** the elements of its $\boldsymbol{z}^b$ innovation vector required by the other PEs.

- Step 9: Calculation of the representer amplitudes by solving a local equation system corresponding to the restriction to area $\boldsymbol{A}$ of the global equations (2). A local representer matrix, $\boldsymbol{R}^a = \boldsymbol{L}^a \left(\boldsymbol{L}^a\right)^T$, and its Hadamard product with a local compactly supported correlation matrix, $\boldsymbol{C}^a \bullet \boldsymbol{R}^a$, are computed. Then, local versions of the m right hand sides of (2b) are calculated as $\boldsymbol{z}^a - \boldsymbol{L}_i + \boldsymbol{e}_i, 1 \le i \le m$. Finally, the local equivalent of (2b) is solved m times, yielding the $\boldsymbol{b}_i$ vectors for the current PE. Since the effective rank of $\boldsymbol{R}^a$ is m rather than n_d^a and as a precaution against $\boldsymbol{R}^a$ losing its positive definiteness due to round off errors, LU decomposition with partial pivoting is used rather than Cholesky decomposition. If LU decomposition fails, singular value decomposition (SVD) is used and near-zero singular values of $\boldsymbol{R}^a + \boldsymbol{W}^a$ are ignored (KR01a).

- Step 10: Computation of the portions of the analysis increments corresponding to each PE-private grid box. A local version of (2c-e) is used. Then, calls to **halo( )** are used to fill the elements of $\boldsymbol{x}^a - \boldsymbol{x}^f$ in the current PE's halo regions. It is more economical to obtain

these elements in this manner than through the application of (2c-e) to each grid box situated within the halo regions.

- Step 11: Transformation of the *T*, *S*, *u* and *v* increments from the analysis levels to averages on the model layers. This step is the reciprocal of step 1. Following this, the analysis increments are inserted gradually into each ensemble member's state vector (Section 3h).

h. Miscellaneous features

1) Incremental analysis updating

Incremental analysis updating (IAU, *e.g.*, Bloom *et al.* 1995) is used to insert the analysis increments, $\boldsymbol{x}^a - \boldsymbol{x}^f$, into the model in a gradual manner. Namely, the model partial differential equations are replaced with

$$\frac{\partial \boldsymbol{x}}{\partial t} = \boldsymbol{F}(\boldsymbol{x}, t) + \frac{(\boldsymbol{x}^a(t_i) - \boldsymbol{x}^f(t_i))}{(t_{i+1} - t_i)}, \quad t_i \le t < t_{i+1}, \qquad (5)$$

where $\boldsymbol{x}^a(t_i)$ and $\boldsymbol{x}^f(t_i)$ are the analysis and forecast at the time, t_i, of the ith analysis.

The IAU is used here for two reasons. First, it lessens the unwanted effects of intermittent data assimilation, specifically initialization shocks resulting from imbalances between the model fields when the analysis increments are inserted directly. Second, it allows the model to gradually adjust the *h* field in response to the *T*, *S*, *u* and *v* increments without violating the constraints imposed by the continuity equation.

2) Measurement operator

In Section 4, the measurement functional, $\boldsymbol{\mathcal{L}}^{a}(\boldsymbol{x})$, is simply a 2D interpolation operator which maps the model temperature field—previously interpolated vertically to a set of levels which include the depths of the TAO measurements—to the latitude and longitude of each observation on the appropriate depth level.

Each PE performs the interpolation to the locations of the observations, $\boldsymbol{d}^{b}$, contained within its PE-private area. Due to the presence of the halo regions, the horizontal interpolation can be made without exchanging information between neighboring PEs. Explicit detail of the interpolation procedure is given in KR01a.

3) Superobservations

As is common when several measurements are made at the same location between successive analyses, the observations are smoothed temporally. This operation, sometimes referred to as superobing and introduced by Lorenc (1981), combines the measurements using weights which decrease exponentially with the time interval between the time of a measurement and that of an analysis in which the measurement is processed. For more detail, the reader is referred to KR01a.

4) Pre-filtering

The purpose of the smoothing operator, Ξ in (1a), is to remove spurious *short-range* covariances from the representer matrix, $\boldsymbol{R}$. These spurious elements result from the limited ensemble size used to estimate the error distribution and from associated sampling errors. Spurious *long-range*

covariances are filtered out by imposing that the covariance functions be compactly supported (Section 3d).

The Ξ operator relies on successive applications of a simple one-dimensional recursive (infinite impulse response) filter which is applied horizontally in each layer to damp small-scale variability prior to calculating $\boldsymbol{L}$ and after subtraction of the ensemble mean from each ensemble member's state vector, as indicated in (1a). The filter equations and response function are discussed in KR01a.

4 Verification

a. Initial test

To test the MvEnKF, TAO temperature data are assimilated every five days into Poseidon for January 1993 to March 1993 using a 40-member ensemble distributed on 256 PEs. For reference, a run without assimilation and one in which the data are assimilated using the UOI are initialized with the initial ocean state corresponding to the MvEnKF central forecast (ensemble member closest to the mean in terms of root mean square distance in the phase space spanned by the model state variables) at the beginning of the experiment.

After the three-month assimilation period, the central forecast from the MvEnKF run is used to initialize a 12-month hindcast run of Poseidon forced with climatological winds, SSTs and heat-

fluxes, and without temperature assimilation. Two similar runs are initialized with the states of the UOI and control runs at the end of the assimilation period.

The purpose of the experiment is merely to verify that the various MvEnKF components are working properly. The assimilation of TAO temperature data into Poseidon and its impact on the model currents, salinity and SSH are the focus of KR01c.

Figure 5 shows the evolution of the spatial-mean temperature anomaly at the TAO-mooring locations, in the observations as well as in the MvEnKF, UOI and control runs. The anomalies shown are with respect to the mean seasonal cycle calculated at each mooring and at each measurement depth for the 1990s. For the MvEnKF run, the anomalies are those of the central forecast.

Initially, the UOI and MvEnKF runs have the same positive bias as the control run, since the same initial ocean state is used in the three runs. After an initial adjustment, the mean anomalies from the MvEnKF and UOI runs are very close to the corresponding observed anomalies during the period with temperature assimilation. In this respect, both methods are effective in correcting the forecast-model bias of the control run. The latter is between 0.5 and 1°C too warm during the initial three months as well as during the following year.

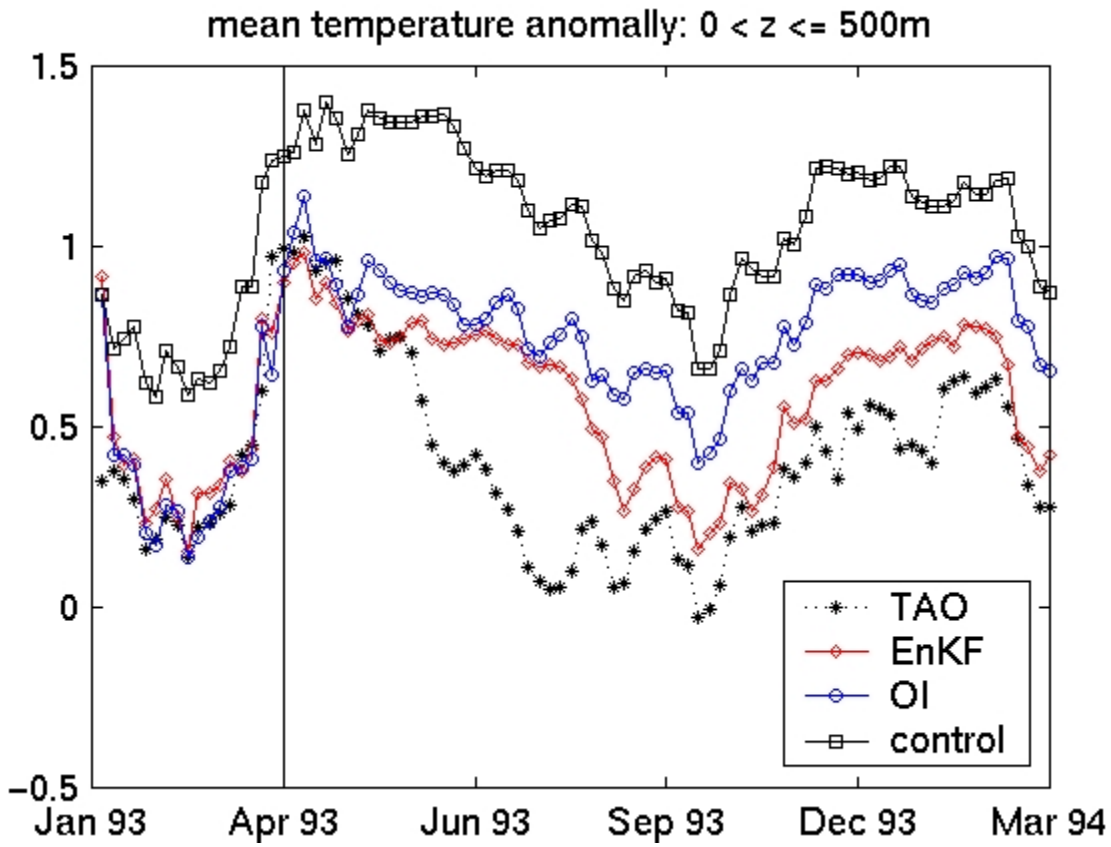

Figure 5. Evolution of spatial-mean temperature anomaly at the TAO-mooring locations during the three-month period with TAO-temperature data assimilation and during the one-year hindcast period without temperature assimilation. The mean anomalies shown correspond to the TAO observations (dotted—stars), the MvEnKF run (solid—diamonds), the UOI run (solid—circles) and the free-model control run (solid—squares).

When the temperature assimilation ceases, the UOI and MvEnKF runs start drifting back towards the warm conditions of the control run. However, even after a year with climatological forcing and no data assimilation, the positive bias of the MvEnKF run (diamonds) is about one third that of the control run (squares). At that point, the level of bias seen in the UOI run (circles) is about two thirds that seen in the control.

It appears from this experiment that (1) the assimilation of the TAO temperature data using either method has a positive impact on the forecast-model bias for temperature and that (2) the better results obtained here with the MvEnKF result in part from the underlying multivariate correction in which not only *T*, but also *S*, *u* and *v*, are updated. Definitive conclusions regarding these two points are drawn in KR01c.

b. Effect of parallel decomposition on analysis

The parallel algorithm relies on the assumption that (1) the analysis calculations can be partitioned resulting in each processor assimilating local data and that (2) the partitioning does not have a deleterious effect on the analysis results. The impact of performing a different local inversion on each processor rather than inverting the global system matrix, $\boldsymbol{S}=\boldsymbol{C}\bullet\boldsymbol{L}\boldsymbol{L}^{T}+\boldsymbol{W}$ in (2b), is examined in this Section.

The local and global solutions are compared through a single temperature analysis using all the TAO mooring data corresponding to January 1, 1997. In this case, sufficiently few data are involved (642 measurements) that (2b) can be solved on each PE without partitioning $\boldsymbol{S}$. Although the number of observations does not necessitate distributing the analysis computations,

the example illustrates how the inversion would be distributed if there were too many data for each PE to process them all at one time, as is the case for the TOPEX altimeter data assimilated in KR01b.

Rather than 40 ensemble members and 256 PEs, as in Section 4a and in KR01b-c, 25 members and 100 PEs are used here, as the corresponding resources suffice for the purpose of this Section.

Figure 6 shows how imposing compact support to $\boldsymbol{R}$ impacts the sparseness of the global $\boldsymbol{S}$. It also illustrates how the sparseness is exploited by distributing the analysis calculations in the parallel algorithm. Figures 7 and 8 illustrate the respective impacts on the assimilation increments of using compactly supported background covariances and distributing the analysis among PEs. As is common, a diagonal $\boldsymbol{W}$ is assumed.

Figure 6a shows the global $\boldsymbol{S}$, when the condition that it be compactly supported is not imposed ($\boldsymbol{L}\boldsymbol{L}^T+\boldsymbol{W}$ in 1c). Figure 7a shows an equatorial section through the corresponding temperature increment. The corresponding sea-surface temperature (SST) increment is shown in Figure 8a.

When the background covariances are compactly supported, the global $\boldsymbol{S}$ ($\boldsymbol{C}\bullet\boldsymbol{L}\boldsymbol{L}^T+\boldsymbol{W}$ in 2b), becomes sparse as Figure 6b illustrates. The most obvious effect of the Hadamard product of $\boldsymbol{C}$ and $\boldsymbol{R}$ on the assimilation increment is that the latter is tapered away from the Equator where no measurements are available (Fig. 8b). The effect of the Hadamard product on the vertical structure of the temperature increment is not as dramatic (Fig. 7b) since the data come from several depths between the surface and 500 meters. The issue of why applying the Hadamard

product, thus solving (2b), is a better idea than solving (1c) is addressed by Houtemaker and Mitchell (2001) in the context of a three-level QG model. They show that the EnKF performs best for small ensembles when the Hadamard product is applied and that the optimal correlation scales are inversely proportional to the ensemble size.

When the analysis is distributed, the calculation of a local $\boldsymbol{S}$ on each PE amounts to sub-sampling the global compactly supported $\boldsymbol{S}$ of Figure 6b. On each PE, the sub-sampling results in a local $\boldsymbol{S}$ which is less sparse than the global $\boldsymbol{S}$ because it does not contain covariances between remote locations which are identically zero as a result of the Hadamard product. Figures 6c-e show local $\boldsymbol{S}$ matrices on three randomly chosen PEs.

Comparing Figure 7c and Figure 7b or Figure 8c and Figure 8b shows that the analysis increments obtained when the analysis calculations are distributed are virtually identical to those obtained with (2), even though the global inversion (2b) is bypassed. Indeed, the root mean square difference between the Equatorial temperature increments of Figures 7b and 7c is 6.0×10^{-4} C. That between the SST increments of Figures 8b and 8c is 1.0×10^{-3} C. Thus, the tremendous computational savings associated with substituting the local $\boldsymbol{S}$ for the global $\boldsymbol{S}$ occur with a negligible impact on the quality of the analysis.

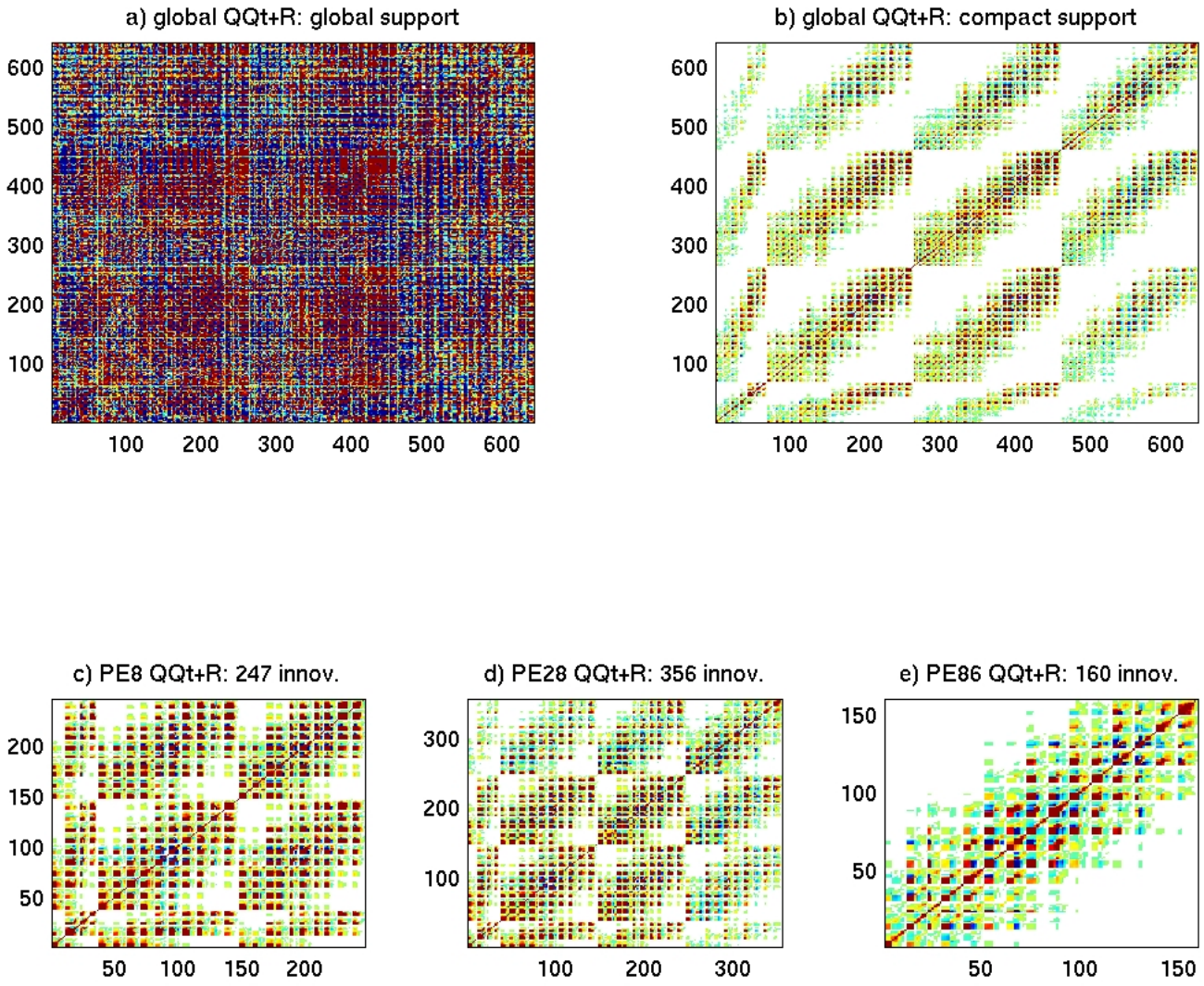

Figure 6. Structure of error-covariance matrices in observation space for one TAO temperature analysis corresponding to January 1, 1997. (a) Global system matrix, ***S***, without compact support. (b) Global compactly supported ***S***. (c-e) Example PE-local ***S*** matrices corresponding to PE 8 for which $n_d = 247$, PE 28 ($n_d = 356$) and PE 86 ($n_d = 160$).

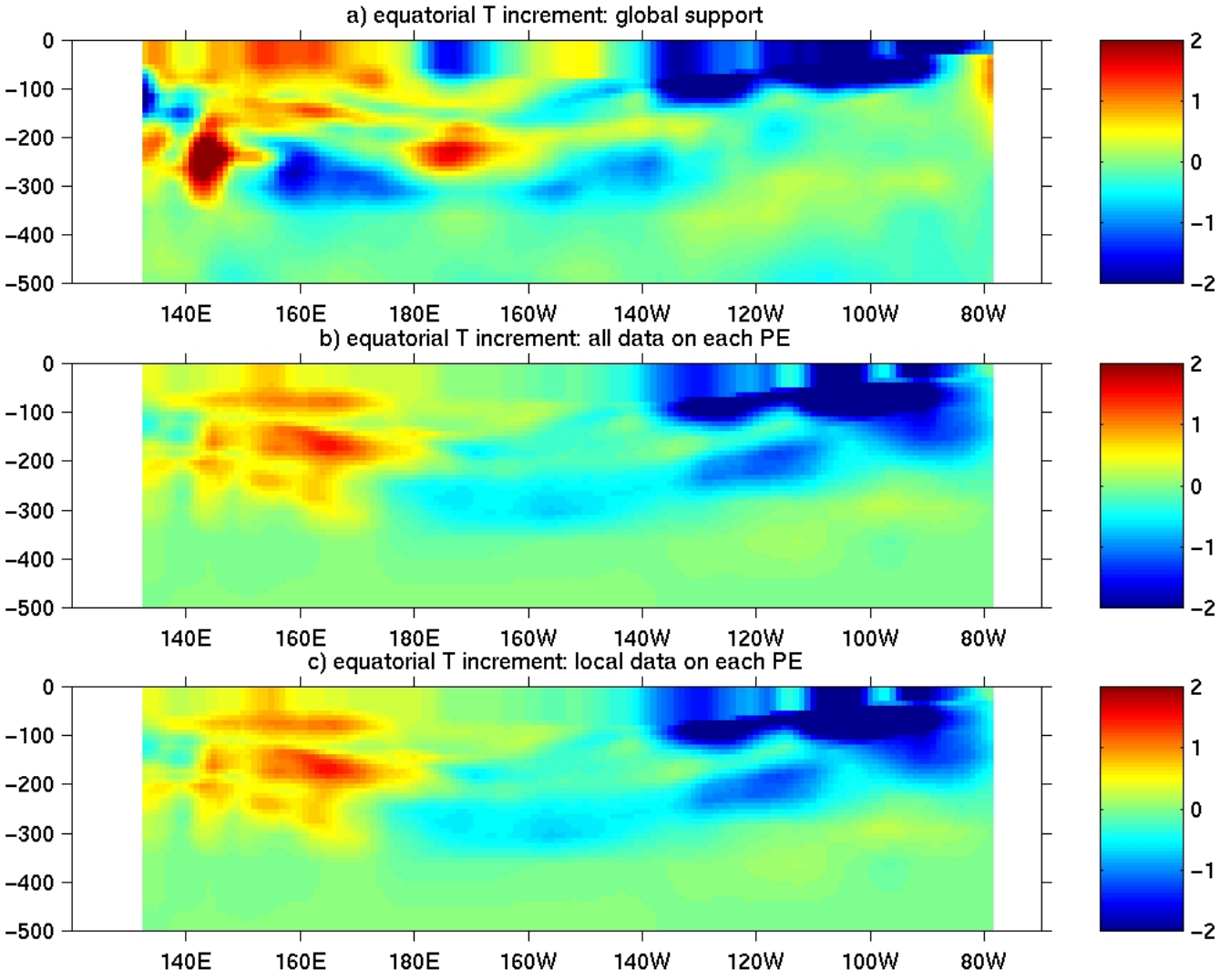

Figure 7. Equatorial sections through the temperature-field part of the analysis increments (°C) obtained after inversion of the matrices shown in Fig.6. (a) Global inversion without compactly supported covariances. (b) Global inversion with compact support. (c) Distributed inversion with compact support.

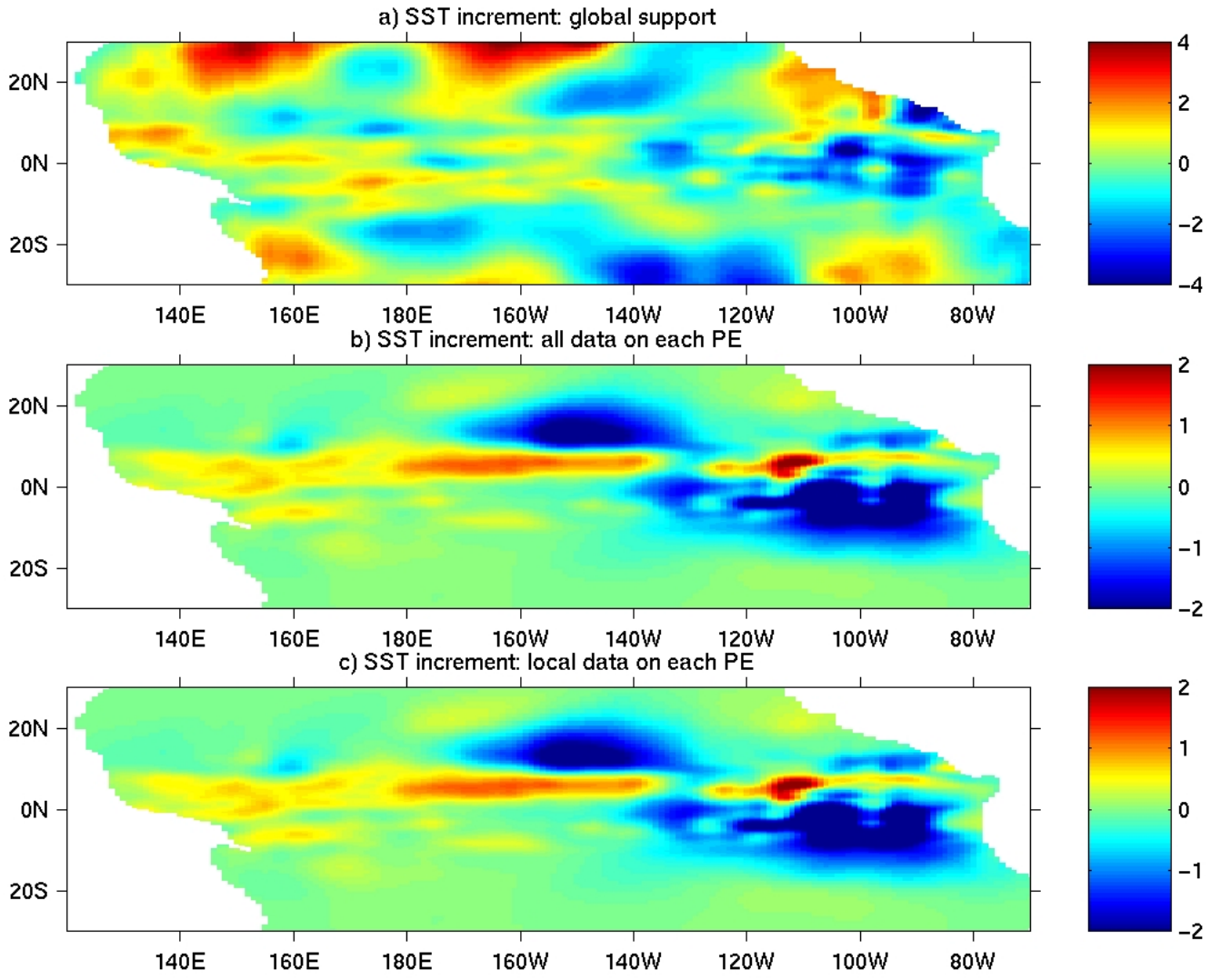

Figure 8. Same as Fig. 7 for the sea-surface temperature increments.

c. Scaling

This Section discusses the two main current limitations of the parallel MvEnKF: (1) that it scales poorly beyond 100 PEs in the present machine (CRAY T3E-600 with 128MB RAM per PE)/model configuration and (2) that the maximum ensemble size attainable is dictated by the memory of the individual PEs on a massively parallel processor (MPP) with distributed memory. Since the expected lifetime of a modern supercomputer is about two years, it is unlikely that these limitations will impose the same restrictions by the time the MvEnKF is used with the global OGCM to initialize the NSIPP production forecasts. Therefore, the software engineering approach used to implement the MvEnKF has focused on portability, modularity and object-oriented design, rather than on optimally using the resources of the current platform.

Figure 9a shows how t_m, the time spent per ensemble member in a five-day analysis cycle involving the assimilation of TAO temperature data, scales with N_{PE} (diamonds). The dashed curve labeled "EnKF perfect" extrapolates the value of t_m for 16 PEs in the range from 16 to 256 PEs, assuming linear scaling. According to Amdhal's law, such scaling can never be achieved. Instead, the time used by an algorithm on p PEs is given by $t_p = t_s(f + (1-f)/p)$, where t_s is the time used by the same algorithm on a serial machine and f is the fraction of the operations that must be performed sequentially.

The observed scaling is not easily compared with theory. First, because t_s is unknown. Second, because f depends on N_{PE}. Still, t_m decreases by a mere 16% when N_{PE} doubles from 128 to 256. Rather, t_m decreases by 45% between 16 PEs and 32 PEs. This is indicative of saturation.

The horizontal resolution of the Pacific basin version of Poseidon used in these experiments is not high enough for the distribution of its state vector over more than 100 PES to be optimal. In contrast, the global version of Poseidon to which the MvEnKF will be applied next has enough state variables to warrant its distribution over more than 100 PEs.. For reference, the observed and perfect scaling curves are also shown for the UOI. In this case, the saturation becomes apparent with 64 PEs at the current model resolution. For each value of N_{PE}, The UOI timing number is higher than the corresponding MvEnKF number because the latter corresponds to the total time divided by the ensemble size (m_{max} below).

In figure 9b, the largest ensemble size allowed by the individual-PE memory on the CRAY T3E-600, m_{max}, is shown as a function of N_{PE}. For each value of N_{PE}, the timing number in Figure 9a corresponds to m_{max} ensemble members, so that memory is saturated. Between 16 PEs and 128 PEs, m_{max} increases approximately linearly from 6 to 36. On 256 PEs, m_{max} is 46.

To increase m_{max} for given N_{PE}, one could simultaneously run several small ensembles on smaller PE partitions rather than a single ensemble on a large partition. However, this would require a communication mechanism not currently supported by the GEMS library. Alternatively, running the MvEnKF on a platform with globally addressable memory would also allow larger ensemble sizes. The 40-member ensembles used in KR01b-c and in Section 4a achieve a good compromise between accuracy and keeping the cost of the data assimilation within acceptable limits.

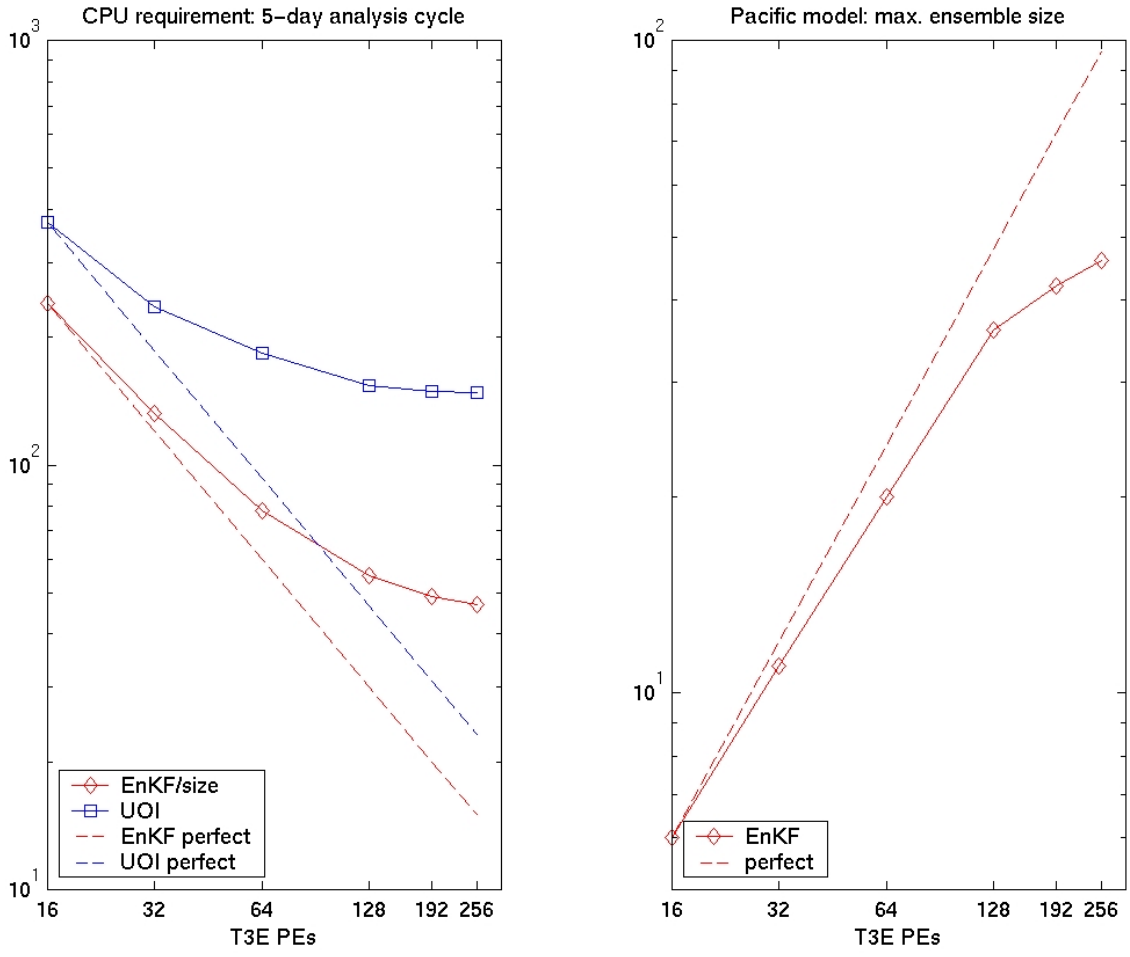

Figure 9. (a) Time *per ensemble member* required to complete one five-day analysis cycle when TAO temperature data are assimilated (t_m in text). The curves labeled "perfect" correspond to an unattainable linear scaling. (b) Largest ensemble size possible as a function of N_{PE} (m_{max} in text) on the CRAY T3E-600.

5 Summary

This article describes the MvEnKF design and its parallel implementation for the Poseidon OGCM. A domain decomposition whereby the memory of each PE contains the portion of every ensemble member's state vector that corresponds to the PE's position on a 2D horizontal lattice is used. The assimilation is parallelized through a localization of the forecast error-covariance matrix. When data become available to assimilate, each PE collects from neighboring PEs the innovations and measurement-functional elements according to the localization strategy. The covariance functions are given compact support by means of a Hadamard product of the background-error covariance matrix with an idealized locally supported correlation function. In EnKF implementations involving low-resolution models, one has the freedom to work with ensemble sizes on the order of hundreds or thousands. Rather, with the state-vector size of approximately two million variables considered here, memory, communications between PEs and operation count limit the ensemble size. In most instances, 40 ensemble members distributed over 256 CRAY T3E PEs are used.

Besides the details of the observing system implementation, the impact of the background-covariance localization on the analysis increments is discussed, as well as performance issues. To confirm that the data assimilation system is working properly, the discussion also includes results from an initial test run in which the MvEnKF is used to assimilate TAO temperature data into Poseidon.

Some issues that must be addressed to improve the MvEnKF are the deficiency of the system-noise model which only accounts for forcing errors, the problem of ensemble initialization which can be addressed using a perturbation-breeding approach, and the memory limitations inherent with running the MvEnKF on a MPP with distributed memory. On a machine with globally addressable memory, the memory-imposed constraints will be less severe. Fortunately, the modular, object oriented approach used to implement the MvEnKF does not tie it to the CRAY T3E architecture.

6 References

Atlas, R., R. Hoffman, S. Bloom, J. Jusem, and J. Ardizzone, 1996: A multiyear global surface wind velocity dataset using SSM/I wind observations. *Bull. Amer. Met. Soc.*, **77**, 869-882.

Bennett, A., 1992: *Inverse Methods in Physical Oceanography*. Cambridge University Press, 346pp.

Blanchet, I., and C. Frankignoul, 1997: A comparison of adaptive Kalman filters for a tropical Pacific Ocean model. *Mon. Wea. Rev.*, **125**, 40-58.

Bloom, S., L. Tacaks, A. DaSilva, and D. Ledvina, 1996: Data assimilation using incremental analysis updates. *Mon. Wea. Rev.*, **124**, 1256-1271.

Borovikov A., and M. Rienecker, 2001: Multivariate error covariance estimates by Monte-Carlo simulation for Pacific Ocean assimilation studies, draft manuscript.

Borovikov, A., M. Rienecker, and P. Schopf, 2001: Surface heat balance in the Equatorial Pacific Ocean: climatology and the warming event of 1994-95. *J. Clim.*, **14**, 2624-2641.

Burgers, G., P. van Leeuwen, and G. Evensen, 1998: Analysis scheme in the ensemble Kalman filter, *Mon. Wea. Rev.*, **126**, 1719-1724.

Cane, M., A.Kaplan, R. Miller, B. Tang, E. Hackert, and A. Busalacchi, 1996: Mapping Tropical Pacific sea level: data assimilation via a reduced state space Kalman filter. *J. Geophys. Res.*, **C101**, 22,599-22,617.

Chen, D., S. Zebiak, A. Busalacchi, and M. Cane, 1995: An improved procedure for El Niño forecasting, *Science*, **269**, 1699-1702.

Daley, R., 1991: *Atmospheric Data Analysis*. Cambridge University Press, 457pp.

Dee, D., and A. Da Silva, 1998: Data assimilation in the presence of forecast bias. *Q. J. R. Meteorol. Soc.*, **124**, 269-295.

Evensen, G., 1994: Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics. *J. Geophys. Res.*, C**99**, 10,143-10,162.

Evensen, G., and P. van Leeuwen, 1996: Assimilation of GEOSAT altimeter data for the Agulhas current using the ensemble Kalman filter with a quasigeostrophic model. *Mon. Wea. Rev.*, **124**, 85-96.

Fukumori, I., and P. Malanotte-Rizzoli, 1995: An approximate Kalman filter for ocean data assimilation - an example with an idealized Gulf-Stream model. *J. Geophys Res.*, **C100**, 6777-6793.

Gaspari, G., and S. Cohn, 1999: Construction of correlation functions in two and three dimensions. *Quart. J. Roy. Meteor. Soc.*, **125**, 723-757.

Gelb, A. (Ed.), 1974: *Applied Optimal Estimation*. MIT Press, 374pp.

Ghil, M., and P. Malanotte-Rizzoli, 1991: Data assimilation in meteorology and oceanography. *Adv. Geophys.*, **33**, 141-266.

Hamill, T., and C. Snyder, 2000: A hybrid ensemble Kalman filter-3D variational analysis scheme, *Mon. Wea. Rev.*, **128**, 2905-2919.

Horn, R., and C. Johnson, 1991: *Topics in Matrix Analysis*. Cambridge University Press, 615pp.

Houtekamer, P., and H. Mitchell, 1998: Data assimilation using an ensemble Kalman filter technique. *Mon. Wea. Rev.*, **126**, 796-811.

Houtekamer, P., and H. Mitchell, 2001: A sequential ensemble Kalman filter for atmospheric data assimilation. *Mon. Wea. Rev.*, **129**, 123-137.

Ji, M., A. Leetmaa, and V. Kousky, 1996: Coupled model forecasts of ENSO during the 1980s and 1990s at the National Meteorological Center, *J. Clim.*, **9**, 3105-3120.

Ji, M., and A. Leetmaa, 1997: Impact of data assimilation on ocean initialization and El Niño prediction, *Mon. Wea. Rev.*, **125**, 742-753.

Kalman, R., 1960: A new approach to linear filtering and prediction problems. *J. Basic Eng.*, **D82**, 35-45.

Keppenne, C., 2000: Data assimilation into a primitive-equation model with a parallel ensemble Kalman filter. *Mon. Wea. Rev.*, **128**, 1971-1981.

Keppenne, C., and M. Rienecker, 2001a: *Design and Implementation of a Parallel Multivariate Ensemble Kalman Filter for the Poseidon Ocean General Circulation Model*, NASA Technical Report Series on Global Modeling and Data Assimilation, Vol. 19, M. Suarez, ed., 31pp.

Keppenne, C., and M. Rienecker, 2001b: Assimilation of satellite altimetry into the Poseidon ocean general circulation model using a parallel ensemble Kalman filter, *J. Marine Sys.*, submitted.

Keppenne, C., and M. Rienecker, 2001c: Validation of a parallel multivariate ensemble Kalman filter implemented for the Poseidon ocean general circulation model., in preparation.

Konchady, M., A. Sood, and P. Schopf, 1998: Implementation and performance evaluation of a parallel ocean model. *Parallel Comput.*, **24**, 181-203.

Lermusiaux, P., and A. Robinson, 1999: Data assimilation via error subspace statistical estimation. Part I: theory and schemes. *Mon. Wea. Rev.*, **127**, 1385-1407.

Lorenc, A., 1981: A global three-dimensional multivariate statistical interpolation scheme, *Mon. Wea. Rev.*, **108**, 701-721.

McPhaden, M., A. Busalacchi, R. Cheney, J. Donguy, K. Gage, D. Halpern, M. Ji, P. Julian, G. Meyers, G. Mitchum, P. Niiler, J. Picaut, R. Reynolds, N. Smith, and K. Takeuchi, 1998: The Tropical Ocean-Global Atmosphere observing system: a decade of progress. *J. Geophys. Res.*, **C103**, 14169-14240.

Mitchell, H., and P. Houtekamer, 2000: An adaptive ensemble Kalman filter. *Mon. Wea. Rev.*, **128**, 416-433.

Niiler, P., and E. Kraus, 1977: One-dimensional models of the upper ocean. *Modeling and Prediction of the Upper Layers of the Ocean*. E. Kraus, Ed., Pergamon, 143-172.

Pacanowski R., and S. Philander, 1981: Parameterization of vertical mixing in numerical models of the tropical oceans. *J. Phys. Oceanogr.*, **11**, 1443-1451.

Pham, D., J. Verron, M. Roubaud, 1998: A singular evolutive extended Kalman filter for data assimilation in oceanography. *J. Marine Sys.*, **16**, 323-340.

Reynolds, R., and T. Smith, 1994: Improved global sea-surface temperature analyses using optimum interpolation. *J. Climate*, **7**, 929-948.

Schaffer, D., and M. Suarez, 1998: Next stop: teraflop; the parallelization of an atmospheric general circulation model. Disponible in Postcript format from `http://nsipp.gsfc.nasa.gov/pubs.html`.

Schopf, P., and A. Loughe, 1995: A reduced-gravity isopycnic ocean model–hindcasts of El-Niño. *Mon. Wea. Rev.*, **123**, 2839-2863.

Troccoli, A., M. Rienecker, C. Keppenne, and G. Johnson, 2001: Temperature data assimilation with salinity corrections: validation in the tropical Pacific Ocean, 1996-1998. *J. Geophys. Res.*, submitted.

Verlaan, M., and A. Heemink, 1997: Tidal flow forecasting using reduced rank square root filters. *Stoch. Hydrol. Hydraul.*, **11**, 349-368.

Yang, S., K. Lau and P. Schopf, 1999: Sensitivity of the tropical Pacific Ocean to precipitation induced freshwater flux, *Clim. Dynam.*, 15, 737-750.

Yuan, D., M. Rienecker, and P. Schopf, 2001: Nonlinear reflection of the equatorial Rossby waves at the Pacific western boundary and its role in ENSO, *J. Clim.*, submitted.